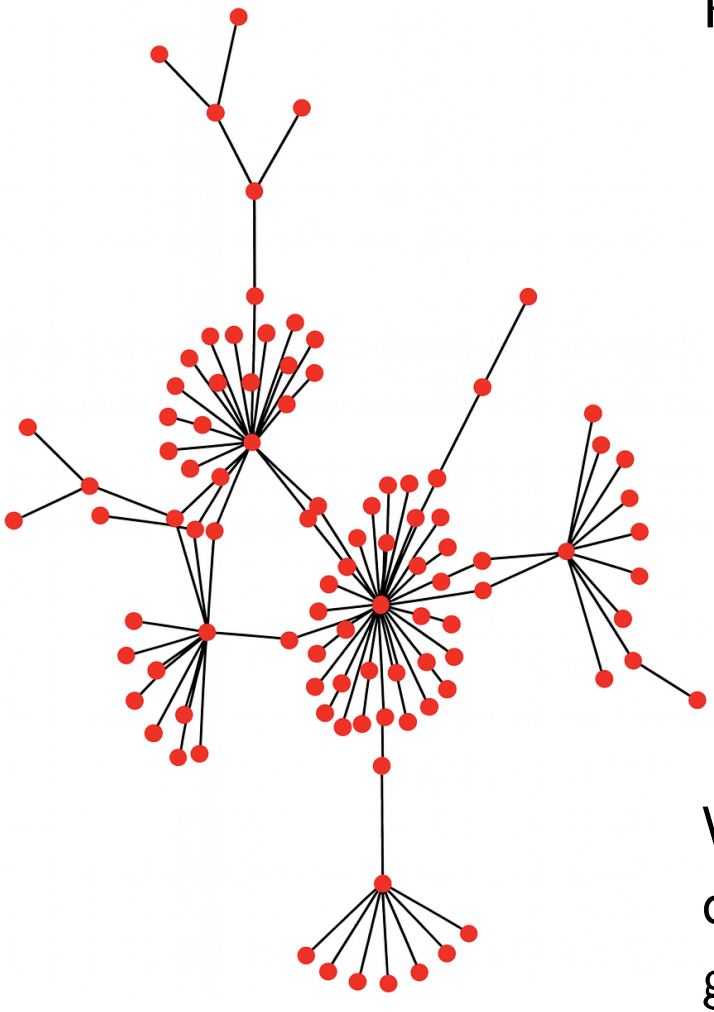
upf. Universitat Pompeu Fabra Barcelona

Seminar 3 · Networks, Crowds and Markets

Scale-Free Networks

Warm-up



Pick the best estimate for:

1. minimal degree:
   0, 1, 3, 30
2. maximal degree:
   5, 25, 100, 1000
3. the diameter:
   2,4,10,20
4. the clustering coefficient:
   0.05, 0.5, 0.75, 0.95.

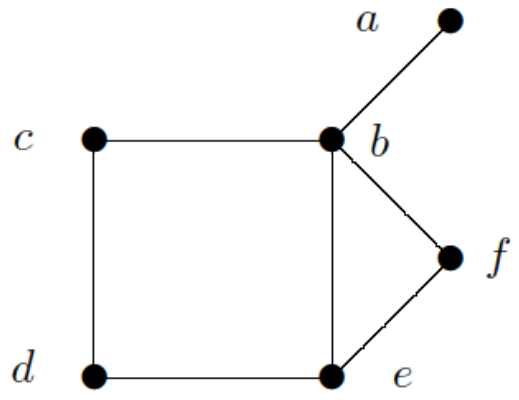
Would a social network be likely to have the diameter and clustering coefficient of this graph?

Exercise 1. Clustering Coefficient

Find the clustering coefficient for:

a) the central node in a Star Graph with N nodes.

b) node b, d and f of the following graph:



Solution to Exercise 1 (sketch)

(a) Central node in a star on N nodes. Let v be the center. Then

$$\deg(v) = N - 1.$$

The neighbors of v are all leaves, and there are *no* edges among them. The local clustering coefficient is

$$C_v = \frac{\#\{\text{edges among neighbors of } v\}}{\binom{\deg(v)}{2}} = \frac{0}{\binom{N-1}{2}} = 0.$$

(b) Nodes b, d, f in the figure. For each such node x:

- list its neighbors $N(x)$ from the picture,
- count how many edges exist inside $N(x)$ (call this E_x),
- then

$$C_x = \frac{E_x}{\binom{\deg(x)}{2}}.$$

Exercise 2. Clustering in ER vs Real Data

Theory. For $G(N, p)$, show $\mathbb{E}[C_v] = p$ for any node v (condition on $\deg(v) = k$, then average over k).

Practice.

- Simulate $G(200, 0.05)$ and estimate the average clustering coefficient $\overline{C}$.
- Compute $\overline{C}$ on a real network (e.g., `karate_club_graph()`).
- Compare. Why does ER under-predict clustering in social networks?

Solution to Exercise 2

Theory. Fix a node v. Condition on $\deg(v) = k$. Among the k neighbors of v, there are $\binom{k}{2}$ possible neighbor–neighbor pairs. In $G(N, p)$, each such edge is present independently with probability p, so

$$\mathbb{E}[\#\text{edges among neighbors of } v \mid \deg(v) = k] = p\binom{k}{2}.$$

The local clustering coefficient is

$$C_v = \frac{\#\text{edges among neighbors of } v}{\binom{k}{2}} \quad (\text{for } k \geq 2),$$

so

$$\mathbb{E}[C_v \mid \deg(v) = k] = p \quad \text{for } k \geq 2.$$

Nodes with degree 0 or 1 are usually defined to have $C_v = 0$ (or omitted); either way they do not change the fact that

$$\mathbb{E}[C_v] = p.$$

Practice. In simulations of $G(200, 0.05)$, the average clustering coefficient $\overline{C}$ will fluctuate around $p = 0.05$.
On a real social-like network (e.g. the karate-club graph), one typically finds $\overline{C} \gg p$ for a comparable edge density. Social networks have *triadic closure*: "friends of my friends" tend to be friends, which an ER model does not encode, so ER strongly under-predicts clustering.

Exercise 3: Estimating clustering coefficient

Suppose you want to compute your clustering coefficient on Facebook or any other social network.

We do not have access to the whole network so we can do things only manually: for any two of your friends check if they are friends.

Say you have 200 friends. The number of pairs – 19,900 – may be too large to explore by hand.

Propose a method to **estimate** the clustering coefficient that can be computed quicker.

Solution to Exercise 3 (clustering sampling)

Clustering coefficient of you (ego node) is

$$C = \frac{\#\{\text{friend–friend edges}\}}{\binom{d}{2}}, \qquad d = \text{number of friends.}$$

Brute force means checking all $\binom{d}{2} = 19{,}900$ pairs.

Sampling idea.

- Uniformly sample M unordered pairs of distinct friends $\{u_i, w_i\}$, $i = 1, \ldots, M$ (e.g. $M = 100$ or $M = 200$).
- For each sampled pair, check manually if they are friends.
- Let $X_i = 1$ if $\{u_i, w_i\}$ are connected, 0 otherwise, and define

$$\widehat{C} = \frac{1}{M} \sum_{i=1}^{M} X_i.$$

Each X_i is a Bernoulli whose expectation is exactly the true probability that a uniformly chosen friend–friend pair forms an edge, i.e. your clustering coefficient. Hence

$$\mathbb{E}[\widehat{C}] = C,$$

so $\widehat{C}$ is an unbiased estimator. Increasing M reduces the sampling variance, but even $M \ll \binom{d}{2}$ is usually enough for a rough estimate.

Exercise 4: Power Law

Given a network with $N = 10^7$ nodes, $k_{min} = 1$, and with the following power-law distribution:

$$p_k = Ck^{-2}$$

a) Determine the probability of finding a node with 100 links attached.

b) What is the expected degree and variance?

c) Determine the value of C in the Continuum Formalism.

d) What is the probability that a node has between 1 and 10 edges?

e) How many hubs do we expect to find in this network? ($k \geq 10^5$)

Solution to Exercise 4: Power law

This could be solved exactly on a computer or approximately using the continuous formalism. We do the latter. Thus we start with (c).

(c) Normalizing constant C. We require

$$1 = \int_1^\infty Ck^{-2}\,dk = C\left[-k^{-1}\right]_1^\infty = C.$$

So $C = 1$, and $p(k) = k^{-2}$, $k \geq 1$.

(a) Probability of degree 100. In the continuum picture, (for the exact answer we would need to use computer)

$$p(k = 100) \approx p(100) = 100^{-2} = \frac{1}{10^4} = 10^{-4}.$$

(b) Expected degree and variance.

$$\mathbb{E}[k] = \int_1^\infty k \cdot k^{-2}\,dk = \int_1^\infty k^{-1}\,dk = \infty,$$

so the mean degree diverges. Similarly

$$\mathbb{E}[k^2] = \int_1^\infty k^2 \cdot k^{-2}\,dk = \int_1^\infty 1\,dk = \infty.$$

(d) Probability $1 \leq k \leq 10$.

$$\mathbb{P}(1 \leq k \leq 10) \approx \int_1^{10} k^{-2}\,dk = \left[-k^{-1}\right]_1^{10} = 1 - \frac{1}{10} = 0.9.$$

(e) Number of hubs with $k \geq 10^5$.

$$\mathbb{P}(k \geq 10^5) \approx \int_{10^5}^\infty k^{-2}\,dk = \left[-k^{-1}\right]_{10^5}^\infty = \frac{1}{10^5} = 10^{-5}.$$

With $N = 10^7$ nodes, expected number of hubs is $N\mathbb{P}(k \geq 10^5) = 10^7 \cdot 10^{-5} = 10^2 \approx 100$ nodes.

Exercise 5: More on power laws

Given a SFN with $N = 25{,}000$ nodes and $\gamma = 2.12$, $k_{min} = 5$, determine:

a) Its degree distribution (in both formalisms).

b) The probability of having a node with exactly 10 links.

c) The expected degree.

d) The number of hubs that the network has (with degree $k \geq 5000$)

e) The expected number of links.

f) What is the probability of finding a node with the same amount or fewer links than the average.

Solution to Exercise 5: More on power laws (1/2)

Let $\gamma = 2.12$, $k_{\min} = 5$.

(a) Degree distribution. *Discrete* (zeta-like) formalism:

$$p_k = \frac{k^{-\gamma}}{\sum_{j=5}^{\infty} j^{-\gamma}}, \quad k = 5, 6, \ldots$$

Continuum formalism: $p(k) = Ck^{-\gamma}, \quad k \geq 5$, with C chosen so that $\int_5^\infty p(k)\,dk = 1$:

$$1 = C\int_5^\infty k^{-\gamma}\,dk = C\left[\frac{k^{1-\gamma}}{1-\gamma}\right]_5^\infty = C\frac{5^{1-\gamma}}{\gamma-1}.$$

Hence $C = (\gamma-1)5^{\gamma-1}$.

(b) Probability of exactly 10 links. Discrete formalism:

$$\mathbb{P}(K=10) = \frac{10^{-\gamma}}{\sum_{j=5}^{\infty} j^{-\gamma}}.$$

(You can approximate numerically if you wish; conceptually this is the answer.)

(c) Expected degree (continuum).

$$\mathbb{E}[k] = \int_5^\infty k\,p(k)\,dk = C\int_5^\infty k^{1-\gamma}\,dk = C\left[\frac{k^{2-\gamma}}{2-\gamma}\right]_5^\infty = C\frac{5^{2-\gamma}}{\gamma-2}.$$

Insert $C = (\gamma-1)5^{\gamma-1}$ to get $\mathbb{E}[k] = \frac{\gamma-1}{\gamma-2}5$. For $\gamma = 2.12$,

$$\frac{\gamma-1}{\gamma-2} \approx \frac{1.12}{0.12} \approx 9.33, \quad \Rightarrow \quad \mathbb{E}[k] \approx 9.33 \times 5 \approx 46.7.$$

Solution to Exercise 5: More on power laws (2/2)

(d) Number of hubs with $k \geq 5000$. Tail probability (continuum):

$$\mathbb{P}(k \geq K_0) = \int_{K_0}^{\infty} Ck^{-\gamma}\, dk = C\frac{K_0^{1-\gamma}}{\gamma - 1} = 5^{\gamma-1}K_0^{1-\gamma} = \left(\frac{5}{K_0}\right)^{\gamma-1}.$$

For $K_0 = 5000$,

$$\mathbb{P}(k \geq 5000) = \left(\frac{5}{5000}\right)^{1.12} = (10^{-3})^{1.12} = 10^{-3.36} \approx 4.4 \times 10^{-4}.$$

Expected number of hubs: $N\mathbb{P}(k \geq 5000) \approx 25{,}000 \times 4.4 \times 10^{-4} \approx 11$.

(e) Expected number of links. Total degree sum is $N\mathbb{E}[k]$, so

$$\mathbb{E}[L] = \frac{N\mathbb{E}[k]}{2} \approx \frac{25{,}000 \times 46.7}{2} \approx \frac{1.17 \times 10^6}{2} \approx 5.8 \times 10^5 \text{ links.}$$

(f) Probability of having $\leq$ average degree. Using the continuum tail:

$$\mathbb{P}(k > \mathbb{E}[k]) \approx \left(\frac{5}{\mathbb{E}[k]}\right)^{\gamma-1} = \left(\frac{5}{(\gamma-1)5/(\gamma-2)}\right)^{\gamma-1} = \left(\frac{\gamma-2}{\gamma-1}\right)^{\gamma-1}.$$

For $\gamma = 2.12$, this is a small number (heavy right tail), so

$$\mathbb{P}(k \leq \mathbb{E}[k]) \approx 1 - \left(\frac{\gamma-2}{\gamma-1}\right)^{\gamma-1}$$

is close to 1. Intuitively: in such a skewed distribution, "average" is pulled up by few hubs, so most nodes have degree below the mean.

Exercise 6: Cayley tree

A *Cayley tree* is a symmetric tree constructed starting from a central node of degree k. Each node at distance 1 from the central node has degree k. More generally, each node at distance ℓ from the central node has degree k until we reach the nodes at distance t, which have degree one and are called *leaves*.

1. Calculate the number of nodes reachable in s steps from the central node.
2. Calculate the degree distribution of the network.
3. Calculate the diameter $d_{\max}$.
4. Find an expression for the diameter $d_{\max}$ in terms of the total number of nodes N.
5. Does the network display the small-world property?

Solution to Exercise 6: Cayley tree (1/2)

(1) Nodes reachable in s steps from the center. Let the center be distance 0.

$$\#\{\text{nodes at distance } 0\} = 1,$$

$$\#\{\text{nodes at distance } \ell\} = k(k-1)^{\ell-1}, \quad \ell = 1, 2, \ldots, t.$$

Hence the number of nodes reachable within s steps ($0 \le s \le t$) is

$$N(s) = 1 + \sum_{\ell=1}^{s} k(k-1)^{\ell-1} = 1 + k\,\frac{(k-1)^s - 1}{(k-1) - 1} = 1 + k\,\frac{(k-1)^s - 1}{k-2} \quad (k \neq 2).$$

(2) Degree distribution. There is:

- 1 node (the center) with degree k;
- for $\ell = 1, \ldots, t-1$, $k(k-1)^{\ell-1}$ nodes with degree k;
- at distance t, leaves of degree 1, number

$$N_{\text{leaf}} = k(k-1)^{t-1}.$$

So all internal nodes ($N - N_{\text{leaf}}$ many) have degree k, and the N_{leaf} nodes have degree 1.

(3) Diameter. The largest distance is between two leaves on "opposite sides" of the tree: leaf $\to$ center $\to$ leaf, which gives $d_{\max} = 2t$.

Solution to Exercise 6: Cayley tree (2/2)

(4) Diameter in terms of N. Total number of nodes (distance $\leq t$):

$$N = 1 + \sum_{\ell=1}^{t} k(k-1)^{\ell-1} = 1 + k\,\frac{(k-1)^t - 1}{k-2}.$$

Solve for $(k-1)^t$:

$$(k-1)^t = 1 + \frac{k-2}{k}(N-1).$$

Thus

$$t = \log_{k-1}\Big(1 + \frac{k-2}{k}(N-1)\Big),$$

and

$$d_{\max} = 2t = 2\,\log_{k-1}\Big(1 + \frac{k-2}{k}(N-1)\Big).$$

For large N, this behaves like $d_{\max} \sim 2\log_{k-1} N$.

(5) Small-world property. Since $d_{\max} = O(\log N)$, the Cayley tree has distances growing logarithmically with N, so it *does* display the small-world property (in the sense of diameter / typical distances).

Additional exercises

Exercise: Power law vs. Poisson

Consider the in-degree distribution of the World Wide Web, which is approximately a power law with exponent $\gamma_{\text{in}} = 2.1$ and minimum degree $k_{\min} = 1$. There are about $N = 10^{12}$ pages.

For comparison, let us take a random Erdős–Rényi network of the same size and with the same average in-degree $\langle k_{\text{in}} \rangle = 4.6$.

1. Estimate the fraction of nodes with $1 \leq k \leq 5$ incoming links in both networks. Compare the results.

2. For the Erdős–Rényi network, find the approximate range of degrees that contains 68% of all nodes.

3. Estimate how many pages have more than 10^5 incoming links in each network, and discuss the qualitative difference.

Hint. For the power law, you may use the normalized form $p_k = k^{-\gamma}/\zeta(\gamma)$. For the ER network, assume a Poisson with $\lambda = 4.6$.

Solution: Power law vs. Poisson (1/4)

We compare two in-degree models on $N = 10^{12}$ pages:

$$\textbf{Power law:} \quad p_k^{\mathrm{PL}} = \frac{k^{-\gamma}}{\zeta(\gamma)}, \quad \gamma = 2.1,\ k \geq 1,\ \zeta(2.1) \approx 1.58.$$

$$\textbf{ER/Poisson:} \quad K \sim \mathrm{Poisson}(\lambda), \quad \lambda = 4.6.$$

Truncation for $k_{\min} = 1$: If we enforce $k \geq 1$ for the Poisson model, use the truncated distribution

$$p_k^{\mathrm{Pois}|k\geq 1} = \frac{\mathbb{P}(K = k)}{\mathbb{P}(K \geq 1)} = \frac{e^{-\lambda}\lambda^k/k!}{1 - e^{-\lambda}}, \quad k \geq 1,$$

where $1 - e^{-\lambda} \approx 0.9900$. This reweights probabilities by a factor $\approx 1/0.99 \approx 1.0101$ (a $\sim$1% effect).

Solution: Power law vs. Poisson (2/4)

(1) Fraction with $1 \le k \le 5$

Power law:

$$\mathbb{P}_{\mathrm{PL}}(1 \le k \le 5) = \frac{\sum_{k=1}^{5} k^{-2.1}}{\zeta(2.1)} \approx \frac{1.422}{1.58} \approx 0.90 \quad \Rightarrow \quad \text{about } 9.0\times10^{11} \text{ nodes.}$$

Poisson ($k \ge 1$ truncated):

$$\mathbb{P}(1 \le K \le 5 \mid K \ge 1) = \frac{\mathbb{P}(1 \le K \le 5)}{1 - e^{-\lambda}} \approx \frac{0.669}{0.990} \approx 0.676.$$

(2) "68%" confidence intervals

For $K \sim \mathrm{Poisson}(\lambda)$, the *normal approximation* gives

$$\mu = \lambda = 4.6, \quad \sigma = \sqrt{\lambda} \approx 2.144.$$

About 68% of a normal law lies in $[\mu - \sigma, \mu + \sigma]$. With a continuity correction:

$$[\,\mu - \sigma,\ \mu + \sigma\,] \approx [\,2.46,\ 6.74\,].$$

The integer degrees in this band are $k \in \{3, 4, 5, 6\}$ (or one may report $\{2, \ldots, 7\}$ for $\approx 68\%$ by symmetry).

Solution: Power Law vs. Poisson (3/4)

(3) Nodes with $k > 10^5$ incoming links. Power law:
For a power-law tail with exponent $\gamma = 2.1$,

$$\mathbb{P}(K > k_0) \approx \frac{1}{\zeta(\gamma)} \int_{k_0}^{\infty} x^{-\gamma}\, dx = \frac{k_0^{1-\gamma}}{\zeta(\gamma)(\gamma - 1)}.$$

With $k_0 = 10^5$ and $\zeta(2.1) \approx 1.58$: $\mathbb{P}(K > 10^5) \approx 1.8 \times 10^{-6}$.
Since $N = 10^{12}$: $N\,\mathbb{P}(K > 10^5) \approx 1.8 \times 10^6$.

Interpretation: A power-law network with 10^{12} nodes contains *millions* of extremely high-degree hubs.

Poisson / ER model: For a Poisson distribution with mean $\lambda = 4.6$, the tail decays *exponentially*. A standard Chernoff estimate gives

$$\mathbb{P}(K \geq k_0) \;\leq\; \exp(-c\, k_0) \quad \text{for some constant } c > 0.$$

For $k_0 = 10^5$, this is roughly $\mathbb{P}(K \geq 10^5) \approx \exp(-10^5) \approx 0$. Thus even with $N = 10^{12}$ pages: $N\,\mathbb{P}(K \geq 10^5) \approx 0$.

Conclusion: Power-law networks naturally generate extremely large hubs; Poisson/ER networks do not.

Solution: Why Poisson Cannot Produce Hubs (4/4)

Poisson tails decay exponentially.
If $K \sim \mathrm{Poisson}(\lambda)$ with $\lambda = 4.6$, then for large k,

$$\mathbb{P}(K = k) \approx \frac{e^{-\lambda}\lambda^k}{k!} \approx \exp\big(-k\log(k/\lambda) + O(k)\big),$$

which is an *extremely* fast decay.

Even multiplying by the total number of pages, $N = 10^{12}$, gives

$$N\,\mathbb{P}(K \geq 10^5) \approx 0.$$

Intuition: A Poisson degree distribution is tightly concentrated around its mean. Huge deviations (like degree 10^5) are essentially impossible. Power-law distributions decay much more slowly, so they *can* produce extreme hubs even in finite networks.

Exercise: Snobbish network

Consider a network of N blue and N red nodes. Any two nodes of the same color are connected independently with probability p, and any two nodes of different colors with probability $q \leq p$.

(a) Compute the expected degree of a blue node:
- within its own color class (blue–blue links),
- and in the full network (blue–blue plus blue–red links).

(b) For large N, what happens to the network when $q = 0$? Describe qualitatively how the picture changes as q increases from 0 to values comparable to p.

(c) Suppose p and q are both of order $1/N$ so that the average degree stays around a constant. Argue (heuristically, not rigorously) that in this regime the typical distance between two nodes still grows roughly like $\log N$, even when $p \gg q$.

Solution: Snobbish network

There are N blue and N red nodes.

(a) Expected degree of a blue node. Fix a blue node B.

- Blue–blue neighbors: there are $N-1$ other blue nodes, each linked with probability p. Contribution

$$\mathbb{E}[\deg_{\text{blue}}(B)] = (N-1)p.$$

- Blue–red neighbors: there are N red nodes, each linked with probability q. Contribution

$$\mathbb{E}[\deg_{\text{red}}(B)] = Nq.$$

Total expected degree of B: $\mathbb{E}[\deg(B)] = (N-1)p + Nq$.

(b) Behaviour as q increases. For large N:

- If $q = 0$, the network splits into two independent ER graphs: one on the N blue nodes, one on the N red nodes, each $\sim G(N, p)$. There is no path between colors.
- As soon as $q > 0$, cross-color edges appear. For moderate q, the two "communities" (blue and red) are still visible but linked by a set of "bridge" edges.
- When q becomes comparable to p, the network looks more homogeneous; color becomes less informative about connectivity.

(c) p, q of order $1/N$: typical distance. Suppose $p = \frac{c_1}{N}, q = \frac{c_2}{N}$, so that expected degree is

$$\mathbb{E}[\deg(B)] \approx c_1 + c_2$$

(a constant w.r.t. N). This is a sparse random graph. Heuristically, from the perspective of a node, the neighborhood grows like a (multi-type) branching process with mean offspring $\approx c_1 + c_2$. As long as the effective branching factor exceeds 1, the size of the ball of radius r grows roughly like $(c_1 + c_2)^r$, so to reach $O(N)$ nodes we need r of order $\log N$. Thus typical distances remain $O(\log N)$ even when $c_1 \gg c_2$ (i.e. $p \gg q$).